Supporting Information

Johnson 10.1073/pnas.0804538105

SI Text

Here, I provide additional information regarding the statistical model used to perform inference regarding the relative merit of R01 proposals as described in the main text and provide details of higher stage assumptions on model hyperparameters, posterior estimation, procedures, and model diagnostics.

Higher Stage Models. For many of the first-stage model hyperparameters, the large volume of data available for parameter estimation obviates the need for careful prior specification. The exception to this rule occurs for the category threshold vectors γ_m, which contain 40 components. These vectors must be estimated not only for large study sections in which proposals are rated by 30 or more reviewers, but must also be estimated in special emphasis panels in which only 3 or 4 raters score a single proposal. An informative second stage prior model for these vectors is thus potentially important for interpreting scores collected from smaller study sections. A prior model for the vectors γ_m was defined as a transformation of Dirichlet probabilities to the $N(0,1 + \sigma_0^2)$ scale upon which reader pre-scores were collected. Letting $p_{m,c}$ denote the probability that a reader in study section m assigns pre-score c to a randomly selected proposal, and letting $C = 41$ denote the number of ordinal categories into which proposals were assigned, then a prior density on γ_m was defined by assuming that

$$\mathrm{p}_m = \{p_{c,m}\}_{c=1} \sim \mathrm{Dir}(\alpha), \qquad \alpha = \{\alpha_\mathrm{h}\}, \qquad [1]$$

where

$$p_{m,c} = \Phi(\gamma_{m,c}\colon 0,\, 1 + \sigma_0^2) - \Phi(\gamma_{m,c-1};\, 0,\, 1 + \sigma_0^2), \qquad [2]$$

$\gamma_{m,0} = -\infty$, $\gamma_{m,c} = \infty$, and $\Phi(\cdot;\, a,\, b)$ denotes the distribution function of a normal random variable with mean a and variance b.

Proper prior densities were assumed for the remaining model parameters. Components of α were assumed *a priori* to be independently distributed according to Cauchy distributions truncated to the interval $(0, \infty)$. Similar results were obtained when the Jeffreys prior for a Dirichlet parameter was assumed for α. The prior densities assumed for the parameters a, b, and c were assumed to be unit exponential densities. These priors reflect a prior belief that scoring weights can assume values close to 0, and such values are most likely when these three hyperparameters assume values <1. Independent inverse gamma distributions with unit scale and shape parameters were assumed for the values of the variance parameters σ_0^2, σ_1^2, σ_2^2 and τ^2. Although long-tailed, this distribution places significant prior mass <1.0 and has its mode at 0.5. Because the latent proposal scores were assumed *a priori* to be independently generated from a standard normal distribution, rater variance parameters >2.0 were regarded as unlikely. Varying the hyperparameters assumed for the distributions of $(a, b, c, \sigma_0^2, \sigma_1^2, \sigma_2^2, \tau^2)$ within a factor of 2 did not lead to significant changes in the posterior distributions on these parameters.

Posterior estimation and model checks. The volume and structure of the proposal rating data and the hierarchical model specification prevent fitting of model parameters with standard software packages. As a consequence, the author created customized C code that implemented a random walk Metropolis–Hastings algorithm to sample from the posterior distribution on the parameter space (e.g., refs. 1–3). An outline of procedures used to validate this code follow.

To begin, an arbitrary subset of the R01 data were selected for study. This subset contained data collected from 20 review groups over two rating cycles for 549 proposals. Attention was restricted to this subset to speed the convergence of MCMC algorithms, which was particularly useful during the model validation phase when numerous variations of the model were fit to data. Final summaries of proposal merit within this subset were based on running the Metropolis-Hastings algorithm for a burn-in period of 100,000 updates; 150,000 subsequent updates were then used for model inference.

Aside from function-level programming checks performed during code development, the final code was validated by using data simulated from the assumed model for a variety of values of model hyperparameters (i.e., σ_0^2, σ_1^2, $\sigma,_2^2$, τ^2, a, b, c, and α). To replicate the data structure, reader pre-scores, reader post-scores and non-reader scores were simulated for scores actually observed in the data. Missing values in the real data were left as missing in the simulated data. The MCMC algorithm was applied to several sets of data simulated in this way, and in each case the correspondence between the posterior distributions of model parameters and data-generating parameters was examined. For reader, proposal, and study section parameters, correlations between simulation truth and posterior means were assessed, whereas the posterior distributions on higher-level model hyperparameters (e.g., variance parameters and Dirichlet mixing parameters a, b and c) were compared with their true values. Satisfactory associations were achieved in all cases. Convergence of the MCMC algorithm was monitored by tracking values of model hyperparameters, and values of $\bar{\mu}$, $\Sigma\mu_i^2$, and Σr_j^2, across iterations.

Simulated data were also used to evaluate whether general features of the data were captured by the hierarchical structure assumed for the generation of reader post-scores and non-reader scores. Figs. S1 and S2 depict histogram displays of non-reader scores for data simulated from the model and actual data. A posterior sample of hyperparameter values $\{\sigma_0^2, \sigma_1^2, \sigma_2^2, \tau^2, a, b, c, \alpha\}$ was used to simulate the data reflected in Fig. S1; all other model parameters and data were generated from these. These figures illustrate the locations of non-reader proposal scores relative to the minimum and maximum reader post-scores as a function of the difference between minimum and maximum post-scores. Subplots in these figures were constructed by identifying all proposals for which the minimum and maximum reader post-scores differed by a specified value, and then constructing a histogram of the difference between the non-reader scores and the minimum reader post-score. In general, when the differences between maximum and minimum reader post-scores is greater than ≈ 0.6, the majority of non-reader scores tends to be distributed approximately uniformly within the range established by the minimum and maximum values. When this range is <0.6, non-reader scores tend to be concentrated near the midpoint of this interval and a slightly higher proportion of scores fall outside of the range defined by the readers. Note the close correspondence between the shapes of the histograms depicted in Figs. S1 and S2. More formal model assessment was performed by comparing the posterior distributions of pivotal quantities to their nominal distributions. Fig. S3 displays a quantile-quantile plot of proposal means observed at the end of a MCMC run of the algorithm. Because the distribution of a pivotal quantity evaluated at a draw from the

posterior distribution is the same as the distribution of a pivotal quantity evaluated at the data generating parameter, these plots can be used to construct model diagnostics with known reference distributions (4). For example, at a value of μ sampled from the posterior distribution, $\bar{\mu} = \Sigma_j \mu_j / J$ is marginally distributed as a $\Phi(0, 1/J)$ random variable, and $s^2 = \Sigma_j \mu_j^2$ is marginally distributed as a χ_J^2 random variable. The observed values of $\bar{\mu}$ and s^2 were -0.10 and 0.87, respectively, for the values displayed in Fig. S3. Both values indicate some model lack-of-fit, which may be partially explained by the tendency of raters to score proposals on 1/2 unit values. It is also likely that the distributions of reader and non-reader errors is not normal on the scale of measurement assumed for the proposal merits. However, these deviations do not appear overly severe and suggest that this baseline model is adequate for obtaining first-order approximations to the ordering of proposal merits with study sections, as well as the uncertainty inherent to these orderings. Posterior means of model hyperparameters appear in Table S1.

1. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys*, 21:1087–1092.
2. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.
3. Gelfand A, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85:398–409.
4. Johnson VE (2007) Bayesian model assessment using pivotal quantities. *Bayesian Anal* 2:719–734.

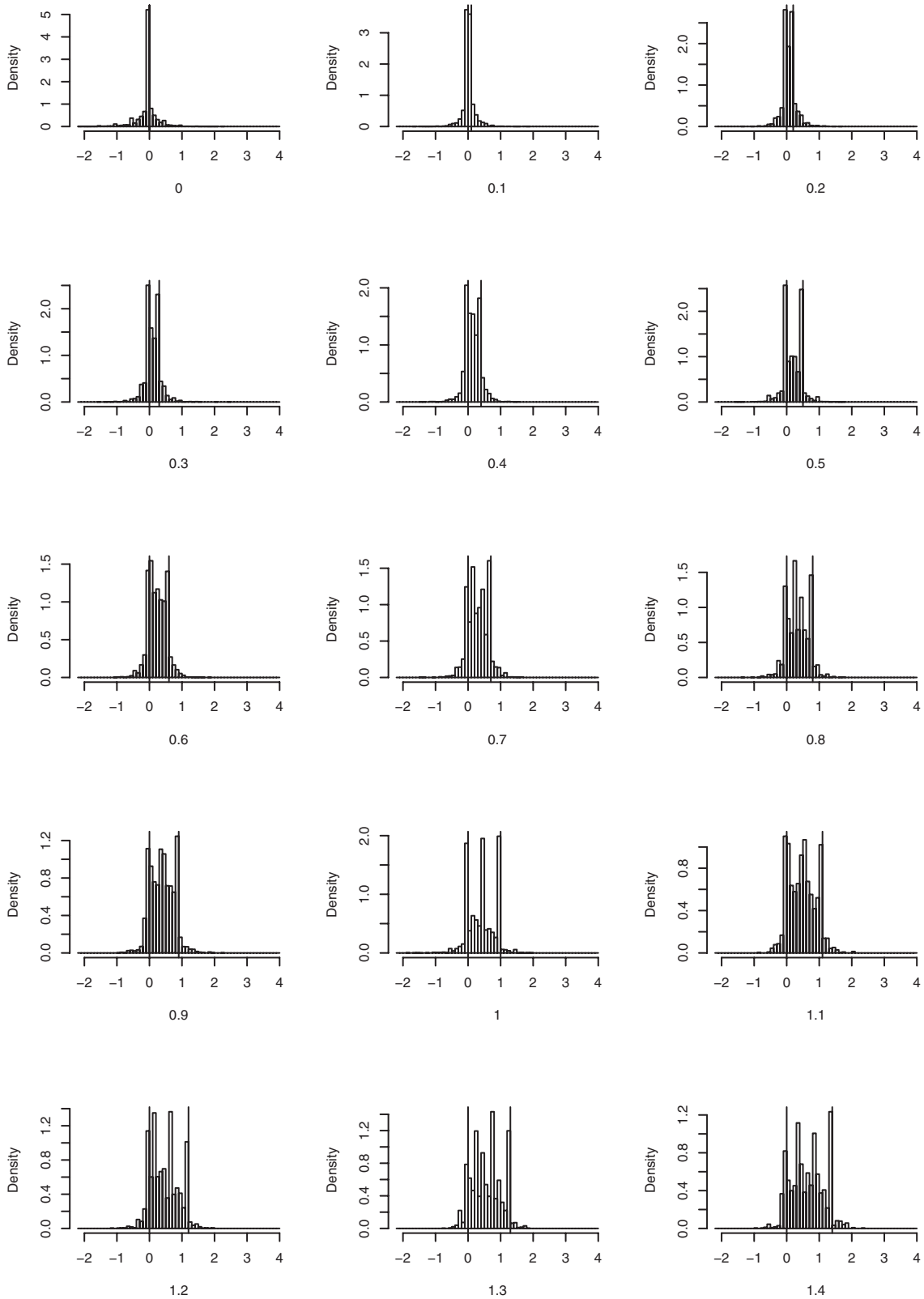

Fig. S1. Plots of the difference between simulated non-reader scores and the minimum simulated reader post-score for various values of difference between the maximum and minimum reader post-scores. The horizontal axes are labeled with latter difference, and vertical lines indicate the interval defined by the simulated reader post-scores.

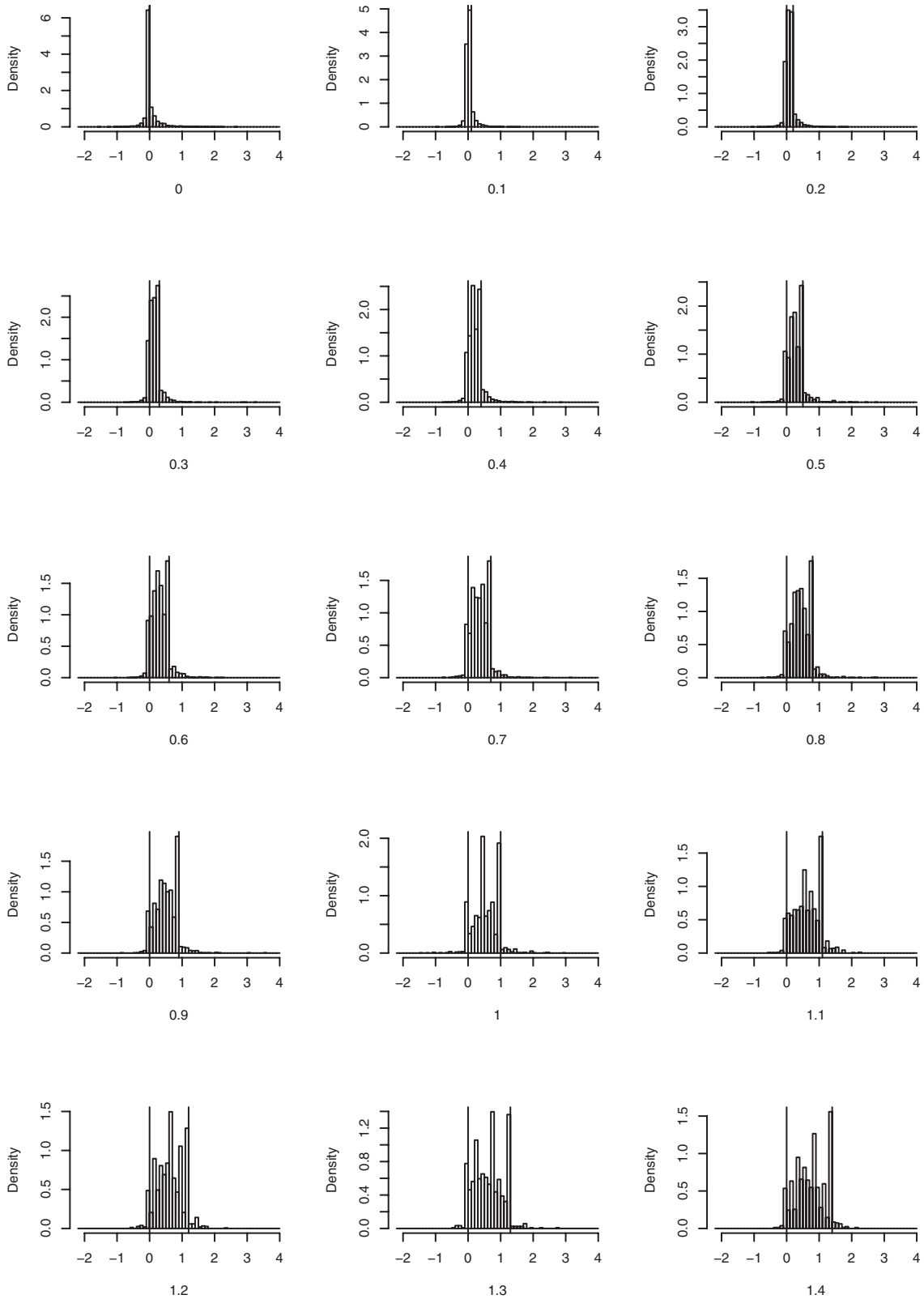

Fig. S2. Plots of the difference between non-reader scores and the minimum reader post-score for various values of difference between the maximum and minimum reader post-scores. The horizontal axes are labeled with latter difference, and vertical lines indicate the interval defined by the reader post-scores.

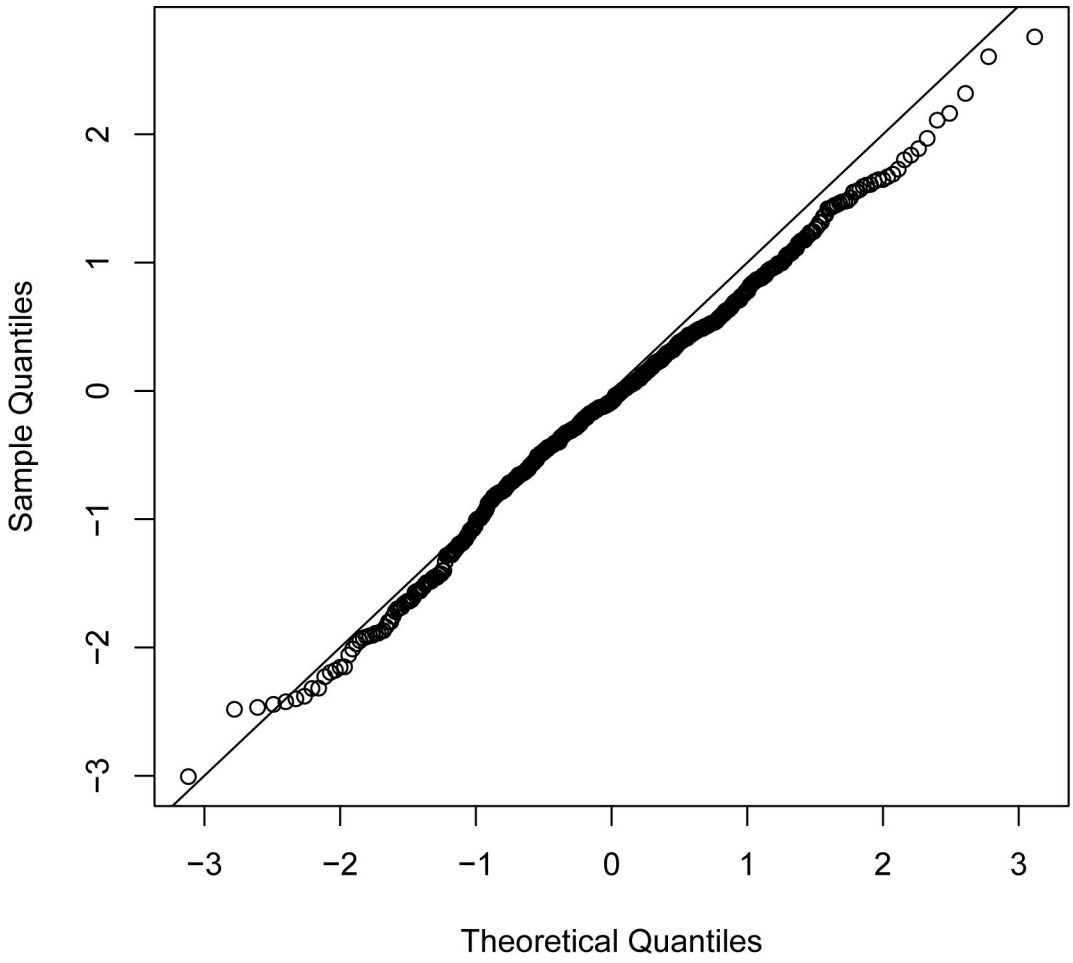

Fig. S3. Normal scores plot of a posterior sample of $\{\mu_i\}$ values.

Table S1. Posterior means of model hyperparameters

Hyperparameter	σ_0^2	σ_1^2	σ_2^2	τ^2	a	b	c	ζ
Posterior Mean	0.41	0.041	0.017	0.041	0.61	0.47	0.19	-.16